Plant Biotechnology 36, 213–222 (2019)
DOI: 10.5511/plantbiotechnology.19.0822a

Original Paper

SIMON: Simple methods for analyzing DNA methylation by targeted bisulfite next-generation sequencing

Simon Vial-Pradel[1], Yoshinori Hasegawa[2], Ayami Nakagawa[1,a], Shido Miyaki[3], Yasunori Machida[4], Shoko Kojima[1], Chiyoko Machida[1,*], Hiro Takahashi[3,5,**]

[1] Graduate School of Bioscience and Biotechnology, Chubu University, Kasugai, Aichi 487-8501, Japan; [2] Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818 Japan; [3] Graduate School of Horticulture, Chiba University, Matsudo 648, Matsudo, Chiba 271-8510, Japan; [4] Division of Biological Science, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8602, Japan; [5] Graduate School of Medical Sciences, Kanazawa University, Kakuma-machi, Kanazawa, Ishikawa 920-1192, Japan
* E-mail: cmachida@isc.chubu.ac.jp Tel & Fax: +81-568-51-6276
** E-mail: takahasi@p.kanazawa-u.ac.jp Tel & Fax: +81-76-234-4484

Received June 28, 2019; accepted August 22, 2019 (Edited by T. Mizoguchi)

Abstract DNA methylation in higher organisms has become an expanding field of study as it often involves the regulation of gene expression. Although Whole Genome Bisulfite Sequencing (WG-BS) based on next-generation sequencing (NGS) is the most versatile method, this is a costly technique that lacks in-depth analytic power. There are no conventional methods based on NGS that enable researchers to easily compare the level of DNA methylation from the practical number of samples handled in the laboratory. Although the targeted BS method based on Sanger sequencing is generally used in this case, it lacks in-depth analytic power. Therefore, we propose a new method that combines the high throughput analytic power of NGS and bioinformatics with the specificity and focus offered by PCR-amplification-based bisulfite sequencing methods. We use in silico size sieving of DNA-fragments and primer matchings instead of whole-fragment alignment in our bioinformatics analyses, and named our method SIMON (Simple Inference for Methylome based On NGS). The results of our targeted BS method based on NGS (SIMON method) show that small variations in DNA methylation patterns can be precisely and efficiently measured at a single nucleotide resolution. SIMON method combines pre-existing techniques to provide a cost-effective technique for in-depth studies that focus on pre-identified loci. It offers significant improvements with regard to workflow and the quality of the acquired DNA methylation information. Because of the high accuracy of the analysis, small variations of DNA methylation levels can be precisely determined even with large numbers of samples and loci.

Key words: bioinformatics, DNA methylation, NGS, sample size calculation, targeted BS sequencing.

Introduction

The field of epigenetics has expanded rapidly over recent years with the development of new techniques, including those for methylome mapping to genomes on the basis of Next Generation Sequencing (NGS) techniques (Wreczycka et al. 2017). Thanks to recent findings, the important role of DNA methylation in the silencing of genes and transposable elements, and its involvement in many biological and cellular processes has been established (Compere and Palmiter 1981; Holliday and Pugh 1975; Phillips 2008). Nevertheless, the function of gene body methylation, comprising CpG methylation in transcribed regions of genes, still remains elusive. Furthermore, the nature of each model organism determines the methods available for its study. For example, until recently, the cost for Whole Genome Bisulfite sequencing of human DNA was prohibitively expensive (Baubec and Akalin 2016); consequently, a large variety of alternative methods were developed (Meissner et al. 2005). The complete methylome of *Arabidopsis thaliana* has been obtained for the wild-type ecotype Col-0 and a few other important mutants, but costs for large scale independent studies of several different samples are still considerable. Techniques based on the sensitivity of restriction enzymes that can

Abbreviations: BS, Bisulfite Sequencing; COBRA, Combined Bisulfite Restriction Analysis; FLASH, Fast Length Adjustment of Short reads; MSP, Methylation Specific PCR; Ms-SNuPE, Methylation-sensitive Single Nucleotide Primer Extension; NGS, Next Generation Sequencing; SIMON, Simple Inference for Methylome based On NGS; Targeted BS-Sanger, Targeted BS method based on Sanger sequencing.

[a] Present address: Institute of Transformative Bio-Molecules, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8602, Japan

This article can be found at http://www.jspcmb.jp/
Published online December 14, 2019

specifically recognize methylated cytosine within their cleavage recognition site were first developed by Bird (1978). This was followed by reports of alternatives developed on the working basis of bisulfite, including Methylation-Specific PCR (MSP), Combined Bisulfite Restriction Analysis (COBRA), Methylation-sensitive Single Nucleotide Primer Extension (Ms-SNuPE), and several other techniques dependent on different applications (Ball et al. 2009; Clark et al. 1994; Cokus et al. 2008; Gonzalgo and Jones 1997; Lister et al. 2008; Rand et al. 2002; Suzuki and Bird 2008; Xiong and Laird 1997). Because studies in mammals have driven the development of many new techniques, they apply better to animal models than to other models, since the density of CpGs, repeat motifs, and restriction sites (e.g., HELP-seq, Suzuki et al. 2010) vary significantly among the organisms of different kingdoms.

One of the methods for understanding the role of gene body methylation is to use a statistical approach and observe how unmethylated, low level methylated, and high level methylated genes behave under different conditions. It would be efficient to do so genome-wide by using enriching techniques. Many genes, however, are not heavily methylated over their entire gene bodies, and could thus easily fall out of the scope of these large scale analyses. The *ETTIN*/*AUXIN RESPONSE FACTOR3* (*ETT*/*ARF3*) gene, for example, is only 13% methylated in the wild-type (Bewick et al. 2016), but this DNA methylation is exclusively concentrated in 3 exons, numbers 6, 9, and 10. We have previously shown that ASYMMETRIC LEAVES1 (AS1) and AS2 of *A. thaliana* are involved in the maintenance of DNA methylation of *ETT*/*ARF3*, while AS1 and AS2 directly repress *ETT*/*ARF3* gene expression (Iwasaki et al. 2013). Our recent results have shown that nucleolar proteins are also positively involved in the maintenance of DNA methylation in *ETT*/*ARF3*. Each of these proteins seems to be differentially involved in the maintenance of several specific combinations of CpG sites. These results imply that the molecular events of CpG methylation are achieved by the differential actions of these proteins (Vial-Pradel et al. 2018). To understand the complex molecular pathways involved in the maintenance of CpG methylation and regulation of gene expression, we need a method that offers single base resolution. Therefore, immune-precipitation, such as Methylated DNA immunoprecipitation sequence (MeDIP-seq) (Weber et al. 2005), and DNA enrichment techniques (Teer et al. 2010) that offer more general information are not suitable for this type of study. To clearly distinguish the patterns of several different samples at once, we also need high coverage similar to ultra-deep sequencing (Jee et al. 2016), which the regular whole genome sequencing technique can only provide at a prohibitive cost.

Generally, one NGS sample contains the converted genomic DNA obtained from one plant-line grown under specific conditions. That is, the converted DNA has had its unmethylated cytosines converted into uracils. By using these samples, it would be possible, although costly, to create libraries for each sample and then perform whole genome sequencing. To obtain enough depth at one given locus, not only must the number of samples be kept low, but a large quantity of DNA is also required in order to sequence the whole genome. Other methods can be used to enrich the CpG sites, for example, by using a restriction enzyme; it is still necessary to consider each sample separately when creating the library, however, and the data processing will have to deal with a large quantity of insignificant DNA fragments. These methods could provide comprehensive information about the methylome, but the cost for processing even one sample is prohibitive. Another alternative is to amplify the loci of interest, and then sequence the PCR products one by one by using conventional sequencing technology. This method does not require state-of-the-art materials and advanced skills, but in order to obtain precise data, several dozens of successfully sequenced molecules are desirable. Since several loci from several different samples are also desired, the amount of work becomes important and multiplication of the steps is a favorable terrain for handling errors.

To simplify the workflow and increase the coverage of our analysis, while targeting only the desired loci for our study, we designed a targeted BS method that can take advantage of both the high throughput power of NGS and the focus offered by PCR amplification. We name our method SIMON (Simple Inference for Methylome based On NGS). Modifications that were made to the PCR step following the bisulfite treatment and data processing have been adapted, accordingly, so as to optimize the computing power and the yield of the analysis, while maintaining high quality standards.

Although comparison with conventional methods is required for novel methods, there has been no conventional method based on NGS for any case whereby researchers try to compare the level of DNA methylation from the practical number of samples handled in the laboratory. Therefore, we compared our SIMON method with the targeted BS method based on Sanger sequencing (targeted BS-Sanger). This comparison shows that the result by targeted BS-Sanger could be reproduced by the SIMON method.

For the conventional methods based on NGS, short reads are generally mapped into reference genomes by using alignments based on all the information for the base sequence of short reads. Alternatively, in the SIMON method, instead of whole-fragment alignment, in silico size sieving of DNA-fragments and primer matchings are used for the assignment of loci, resulting in the optimized usage of computing power.

Materials and methods

Plant materials and growth conditions

A. thaliana ecotype Col-0 (CS1092), *as2-1* (CS3117) and *as1-1* (CS3374) were obtained from the Arabidopsis Biological Resource Center (Columbus, OH, USA; ABRC). We outcrossed *as2-1* with Col-0 three times and used the progeny for our experiments (Kojima et al. 2011). Details of *top1α-1*, *fas2-2*, *rh10-1, as2-1 rh10-1* were described previously (Ishibashi et al. 2013; Matsumura et al. 2016; Takahashi et al. 2002). The *top1α-1 as2-1* and *top1α-1 as1-1* double mutants were generated by crossing each single mutant. The *as1-1*, *rh10-1*, *top1α-1*, and *fas2-2* mutants were on the Col-0 (WT) background. Seeds were first sown on soil, and then after 2 days at 4°C in darkness, plants were transferred to a regimen of white light at 50 μmol $m^{-2} s^{-1}$ for 16 h and darkness for 8 h daily at 22°C for the mutant lines with the *top1α-1* and *fas2-2* mutations, as described previously (Semiarti et al. 2001). The Col-0 ecotype and the mutant lines with *rh10-1* mutations were grown on soil at 26°C.

Genomic DNA extraction

DNA was extracted from about 100 mg of the whole aerial part of 14-day-old plant seedlings. The plant samples were frozen in liquid nitrogen and then crushed into powder in a mortar. Total DNA was isolated with a DNeasy Plant Mini Kit (QIAGEN, Valencia, CA), according to the manufacturer's instructions.

Bisulfite treatment and sample preparation for sequencing

Approximately 300 ng of genomic DNA was used for bisulfite conversion with the EZ DNA Methylation-Gold kit (Zymo Research, Irvine, CA, USA). Immediately after the conversion, we amplified the fragments of interest by using the Epitaq bisulfite kit (TAKARA BIO INC, Kusatsu, Japan). The PCR reactions were carried out with different sets of primers for each DNA sample. Primer sets are listed in Supplementary Table S1. A sequence of 4 different nucleotides, hereinafter referred to as "barcode" was added to the 5′ end of the primers in order to identify the original DNA sample sequences during subsequent sequencing. Therefore, apart from Illumina's i7 index and i5 index, a "sample barcode" of our design (consisting of 4 bases) is located in "insert" which is the internal region between the indexing adapters. We targeted *ETT*/*ARF3* exon 6, *ETT*/*ARF3* exon 9, *ETT*/*ARF3* exon 10, the *APETALA1* (*AP1*) promoter region, *ARF4* exon 10, *PHABULOSA* (*PHB*) exon 14, and *BREVIPEDICELLUS* (*BP*)/*KNAT1* exon 4. The primers were designed to specifically target the coding strand (Iwasaki et al. 2013; Supplementary Table S1). The conversion of each DNA sample was tested for completeness by using the *AP1* promoter, a region previously shown to be non-methylated (*AP1*, AT1G69120). A solution containing approximately the same amount of each fragment from each DNA sample was prepared and purified with the WIZARD SV Gel and PCR Clean-up System (Promega Corporation, Madison, USA) to eliminate small DNA fragments, according to the manufacturer's instructions. The purified solution was then adjusted to a concentration of 200 ng/μl of DNA and a volume of 20 μl.

NGS sequencing

The KAPA HyperPlus Library Preparation Kit (Kapa Biosystems, Wilmington, USA) was used to prepare a sequencing library. DNA (500 ng) was subjected to the End Repair and A-Tailing reaction, according to the KAPA HyperPlus Library Preparation Kit specification. The DNA ligase mix, including annealed adapter, was added to the A-tailed library. The library products were purified with AMPure XP beads (Beckman Coulter, CA, USA) to remove adapter dimers. After a 4-cycle PCR amplification, the library products were purified with AMPure XP beads. Libraries with a 5% PhiX control were sequenced to balance the overall lack of base diversity on an Illumina MiSeq system with a MiSeq Reagent Kit v3 (Illumina Inc., San Diego, USA) generating 2×300 bp paired-end sequences. The output data generated by NGS in this paper have been submitted to the DDBJ Sequence Read Archive (DRA) under the accession number DRA008463.

Data processing

We used the Fast Length Adjustment of Short reads (FLASH) proposed by Magoč and Salzberg (2011) to combine paired reads. FLASH has two important parameters, m: the minimum overlap length, and M: the maximum overlap length. In this study, parameter m is set to 70 and M is set to 240, according to the expected sizes of the PCR products. First, 9,737,840 sequenced read pairs were combined by FLASH, and then 8,593,494 combined reads were constructed. We used in silico size sieving of the DNA-fragments and primer matchings instead of whole-fragment alignment for our bioinformatics analyses, in order to filter the combined reads based on the expected sizes for each PCR product. Then, we determined the direction of each filtered sequence on the basis of pairs of PCR primers, and that of the classified sequences on the basis of each pair of the 4 bp barcodes, which we designed. Following these processes, 4,928,137 sequences remained, from which we selected 3,885,305 sequences with a sequence error rate $<1\%$ and used them for counting methylation events. All programs, except for FLASH, were written in R (www.r-project.org). We also used R libraries, sangerseqR (Hill et al. 2014), CrispRVariants (Lindsay et al. 2016), ShortRead (Morgan et al. 2009), Biostrings and seqinr (Charif and Lobry 2007).

Statistical power analysis

We used an R library, statmod (Bonnefoix et al. 2001) to estimate necessary sample sizes.

Results

Barcode-extended primers for the SIMON method

As shown in Figure 1, our idea was to use modified PCR primers that would allow us to amplify targeted loci so as to identify our sample without the necessity of generating

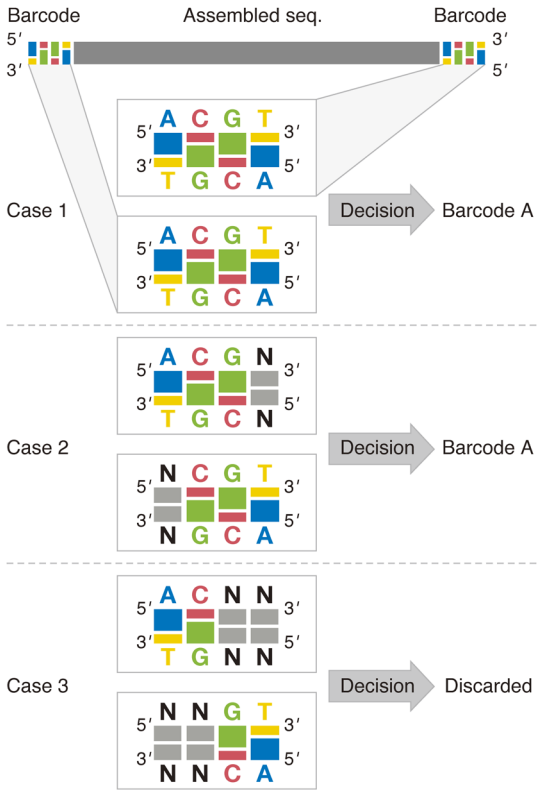

Figure 1. Barcode-extended primers for targeted NGS. Case 1) It is possible to identify unique barcodes in the case that no error exists in the barcode sequences. Case 2) It is possible to identify unique barcodes in the case that one ambiguous nucleotide exists in the barcode sequences. Ambiguous nucleotides at the 5′ end of reads are frequently observed for Illumina NGS in cases where sequence diversity is low; for example, the sequencing for AT-rich samples. Case 3) It is impossible to identify unique barcodes in cases where there are 1< ambiguous nucleotides or 0< substitution errors in barcode sequences.

several libraries. In order to do so, we extended our usual primers for bisulfite-treated DNA with a 4 base-pair (bp) sequence at the 5′ end. This 4 bp sequence is called a sample barcode, and its design relies on simple concepts. It necessarily includes the 4 nucleotides, A, C, G, and T, and each of the 24 combinations of those 4 nucleotides can be used to design primers (Supplementary Table S1). With 4 different nucleotides, if one cannot be detected during sequencing, as is common with Illumina NGS for sequences with low diversity, the 4th can be deduced from the 3 detected nucleotides. This allows more sequences to be used to generate the final data. Studies of DNA methylation should be applied to particular biological aspects, limited only to several loci. By using this method, it is possible to design as many barcode-extended primer pairs as desired. The only limitation is the cost of these oligomers; as such, we examined 7 loci. Whole genome analysis provides important preliminary data to determine what regions of the DNA to study. By using the barcode-extended primers, it should be possible to examine 24 different samples at once, including many mutant and transgenic lines, under different growth conditions. In our situation, we examined 8 different samples. Note that by increasing the number of samples from 8 to 24, the number of hits per sample would be decreased by two-thirds.

Experimental work flow for the SIMON method

As shown in Figure 2A and 2B, after bisulfite treatment of purified DNA, PCR is carried out by using the barcode-extended primers. A barcode is attached to each sample and the barcode-extended primers are used accordingly. Consequently, the initial 8 samples of converted genomic DNA become distributed into 56 different PCR tubes containing the amplified DNA of 7 different loci plus a unique 4 bp sequence at each end of each DNA fragment that is specific to the sample of origin. The workflow is greatly simplified, because all these different DNA fragments can be mixed and sequenced together without losing any information (Figure 2B). This mixture of all DNA fragments from all samples is purified to remove the small DNA fragments, while the 300–600 bp double stranded DNA molecules remain (Figure 2C). It is still necessary to use the Illumina kit to create the library for NGS, but there is no cost for additional samples, because all of the sample information is already contained in the DNA sequence.

Data processing steps of the SIMON method

We propose new changes in the data processing method to take advantage of the barcode sequences and the high precision of the PCR (Figures 2D, 3). The experimental work flow for data processing is shown in Figure 2D, and our experimental results of each data processing step are shown in Figure 3. The first step of the analysis, based on FLASH, combines paired reads: 8,593,494 pairs (88.25% of the detected raw read pairs) could be combined into singleton sequences (DNA fragments), as shown in Figure 3A. Next, we greatly simplified the analysis, since by using a variant of targeted bisulfite sequencing, we know that the DNA fragment sizes after PCR are the length of the DNA sequences between the primers, which depends on each locus plus the barcode bases, which is always 8 bp. By eliminating the combined pairs that did not have any of the expected sizes, we obtained 5,845,303 combined pairs with the expected sizes (68.02% of the combined pairs) for further analysis (Figure 3B). As shown in Figure 4, the combined pairs were distributed around clear peaks at the expected sizes. Numerous DNA fragments that are 1 bp shorter than the expected sizes are likely DNA fragments with a deletion of 1 bp. The origin of these fragments is possibly mistakes during PCR and sequencing, which are more common with converted DNA that has become AT-rich, than with the original genomic DNA.

As shown in Figure 3C, after the selection by size, the sequence of each fragment is examined for identification.

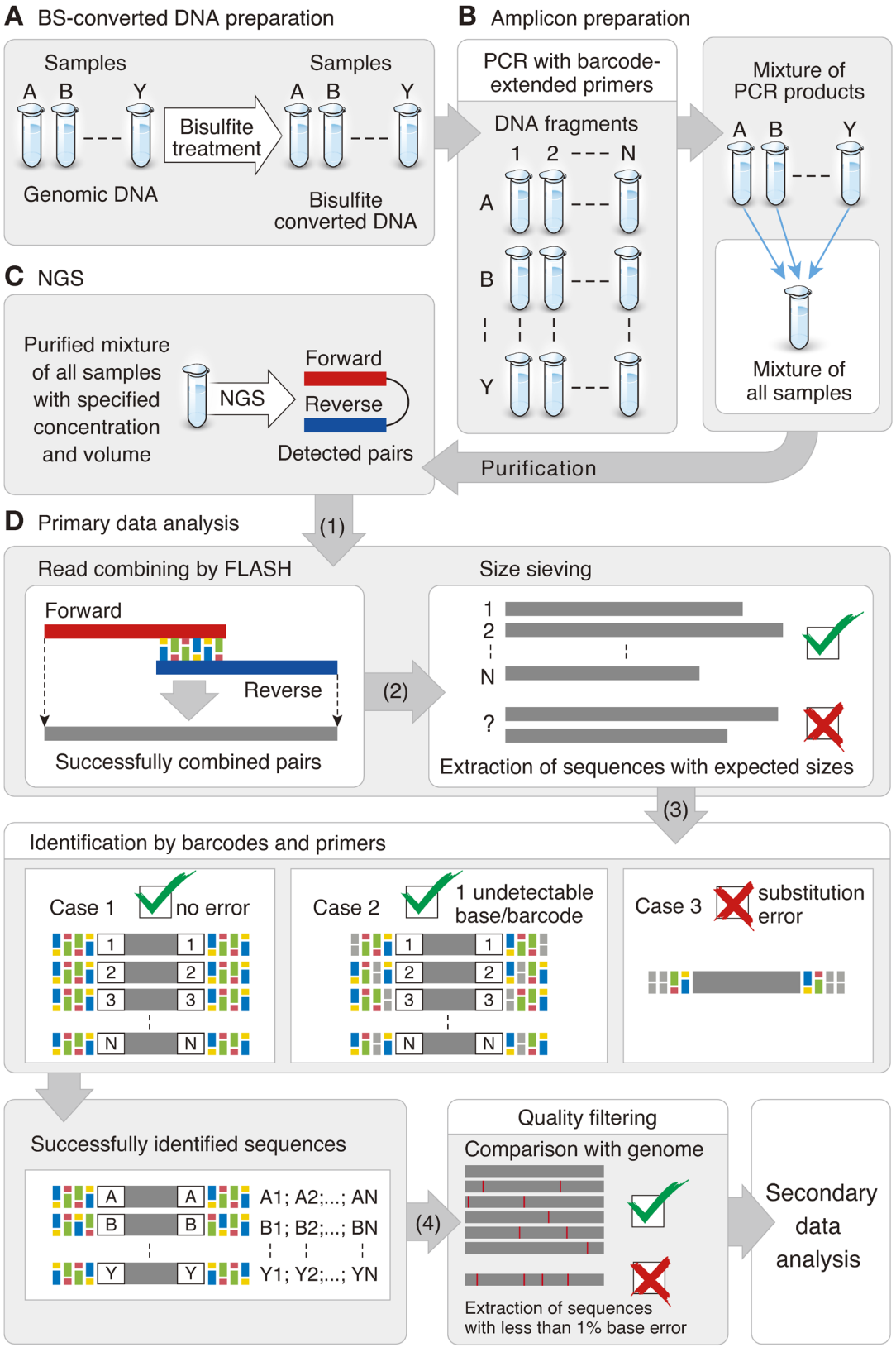

Figure 2. Experimental workflow and data processing of the SIMON method. A) Schematics of the steps for the preparation of DNA samples from plant material and bisulfite treatment, B) schematics of the steps for PCR amplification of loci of interest by using barcode-extended primers and a mixture of the amplicons into one solution, C) schematics of the sample preparation steps required for the NGS run and NGS, and D) schematics of the steps constituting the primary data analysis of the NGS raw data. Each experimental procedure and each analytical step is indicated by an arrow. The details of each step are illustrated or described. The method explains how sample materials are prepared for NGS, and then how the raw data from NGS are processed to provide methylation counts at each cytosine position.

A Read combining by FLASH

Total number of detected pairs 9,737,840 (1) Number of successfully combined pairs 8,593,494

Recovery rate: 88.25%

B Size sieving (2)

Number of counts at expected sizes 5,845,303

Number of hits at each expected sizes	363+8 bp *ETT exon 9*	400+8 bp *BP exon 4*	418+8 bp *ETT exon 10*	419+8 bp *ETT exon 6*	463+8 bp *PHB exon 14*	485+8 bp *AP1 promoter*	504+8 bp *ARF4 exon 10*
	1,249,250 hits	731,307 hits	977,520 hits	1,008,908 hits	921,959 hits	661,152 hits	295,207 hits

Recovery rate: 68.02%

C Identification by barcodes and primers (3)

Number of successfully identified sequences 4,928,137

Number of sequences	*ETT exon 9*	*BP exon 4*	*ETT exon 10*	*ETT exon 6*	*PHB exon 14*	*AP1 promoter*	*ARF4 exon 10*
Barcode A	92,861	61,761	60,253	71,689	56,105	42,577	19,898
Barcode B	125,123	67,852	95,575	92,153	95,034	63,636	21,078
Barcode C	157,610	117,893	86,043	121,865	113,377	75,811	48,663
Barcode D	208,270	137,387	132,615	144,442	175,370	113,212	96,539
Barcode E	197,301	112,120	134,130	150,574	162,493	116,862	53,410
Barcode F	98,460	60,941	93,051	97,410	68,388	75,126	7,864
Barcode G	116,801	53,517	65,265	76,577	86,564	49,301	16,226
Barcode H	82,880	24,012	58,186	62,906	60,788	40,554	11,738

Recovery rate: 84.31%

D Quality filtering and final data (4)

Number of sequences with less than 1% base error 3,885,305

Number of sequences	*ETT exon 9*	*BP exon 4*	*ETT exon 10*	*ETT exon 6*	*PHB exon 14*	*AP1 promoter*	*ARF4 exon 10*
Barcode A	83,743	57,754	49,142	59,420	44,760	22,412	10,149
Barcode B	112,905	63,568	79,957	76,957	75,472	34,096	10,195
Barcode C	140,384	109,358	68,673	98,918	87,150	35,900	21,795
Barcode D	185,204	127,622	105,203	118,733	136,071	55,029	44,673
Barcode E	176,737	104,414	107,525	125,120	127,492	63,664	23,700
Barcode F	87,219	56,637	74,392	78,614	53,396	33,519	3,698
Barcode G	105,408	50,222	53,981	63,438	69,303	26,677	8,664
Barcode H	75,168	22,490	48,549	52,987	48,660	22,456	5,933

Recovery rate: 78.84%

Figure 3. Count of reads and yield of each data processing step. A) The total number of detected sequences and successfully combined sequences, B) the number of sequences of expected sizes and their distribution, C) the number of successfully identified sequences, and D) the distribution and the number of sequences that fulfil the quality criterion and their details, are all shown alongside the yield for each step. The total yield is 39.90% from detection until quality check.

At each size corresponding to a specific pair of primers, these sequences are confirmed in order to eliminate any DNA sequence that does not have the correct primer pairs for its size. As one example, *ETT* exon 6 fragments that have lost 1 bp during PCR or during sequencing, would be only 426 bp long (instead of 427 bp) (Figure 4). Now, 426 bp was the expected size of *ETT* exon 10 DNA fragments, therefore the shortened *ETT* exon 6 DNAs were not eliminated during the previous step. At this step, however, because the specific primers of *ETT* exon 10 were absent in these 426-bp fragments, the program eliminates these shortened fragments. After confirming that the DNA sequences match the DNA sizes, the program examines the barcode sequence in order to identify the sample of each DNA sequence (Figures 1, 2D). There are 2 cases that allow the identification of the sample. In case 1, the 4-nucleotide sequences at the 5′ end of each DNA strand are identical and correspond

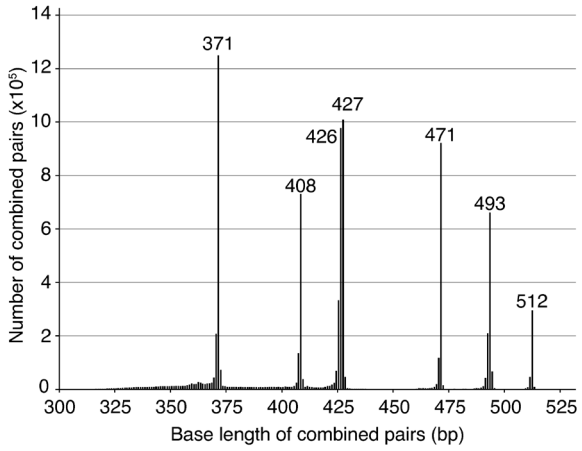

Figure 4. Distribution by size of successfully combined pairs. The expected sizes corresponding to the PCR products are indicated in the horizontal axis, and the numbers of combined pairs for each size are indicated in the vertical axis.

to the sequence of one of the barcodes utilized in the experiment. Identification is then trivial. In case 2, one of the 4 nucleotides in either or both of the 4-nucleotide sequences at the 5′ end of each DNA strand cannot be detected. It is then possible to identify what the missing nucleotide is, when the three others are known, if the resulting sequence corresponds to the sequence of one of the barcodes utilized in the experiment. When DNA sequences with low diversity were sequenced by using Illumina NGS, ambiguous nucleotides having the IUPAC nucleotide code "N" were frequently observed at the 5′ end of reads, but owing to the design of the barcode sequences, so long as no more than one nucleotide is undetectable, we can still identify the sample of origin. It is sometimes preferable to not identify the barcode sequence, or it may simply be impossible to do so. It becomes impossible when too many nucleotides are undetectable simultaneously in both reads, as well as when nucleotide substitution has occurred. It is theoretically possible to use only one barcode to identify the sample, but that increases the chance of incorrect identification. As a precaution, we only used DNA sequences for which the same barcode sequence could be identified in both strands, independently. Following these two identification steps, a total of 4,928,137 sequences (84.31%) divided into 56 groups remained (Figure 3C). The number of sequences appear to be biased in Figure 3D. Those of barcode D and E are more than others and those of barcode F and H are less. These biases might depend on efficiency of PCR caused by amount and/or purity of genomic DNA.

The final step, illustrated in Figure 2D, sets a quality criterion for the sequences to determine which one can be used to count methylation. We decided to use only sequences that have less than 1% base error between each primer, including the intronic regions when present (Figure 3D). The sequences had many different sizes, the shortest one, *ETT* exon 9, has 299 bp between each primer and the longest one, *ARF4* exon 10, has 461 bp. This means that sequences with 3 base errors or more in *ETT* exon 9 will be discarded, whereas sequences with up to 4 base errors in *ARF4* exon 10 will still be extracted. Base errors can consist of ambiguous nucleotides, substitutions, additions, or deletions. Only the sequence of the coding strand is aligned to the genomic sequence and examined. In this case, any substitution at the position of an A, G, or T nucleotide will be counted as an error, while at the C nucleotide position in the genomic sequence, a substitution by a T is not counted as an error, since it means that this particular C was converted, therefore, that it was not methylated. According to this rule, a total of 3,885,305 sequences (78.84%) divided into 56 groups remained (Figure 3D).

In the end, 39.90% of the detected pairs were successfully extracted for the methylation analysis (Figure 3). This method provided an excellent depth of analysis with numbers of sequences per locus per sample ranging from 3,698 sequences (*ARF4* exon 10, barcode F) to 185,204 sequences (*ETT*/*ARF3* exon 9, barcode D), as shown in Figure 3. The amount of work necessary to sequence 4,000 molecules by conventional means is extremely important, and it would represent only the smallest portions of the data provided by our method. Similarly, in order to obtain a similar coverage with Whole Genome Sequencing, a considerable amount of repetition would be necessary, which would mean considerable expense.

Discussion

The SIMON method is an advanced method based on NGS for the targeted BS-Sanger. Although novel methods need to be compared with the conventional methods, there is no conventional method based on NGS for a case where the researchers attempt to compare the level of DNA methylation from the practical number of samples handled in the laboratory. Therefore, we compared the SIMON method and the targeted BS-Sanger for methylation levels in all cytosine positions of the coding strand of *ETT*/*ARF3* exon 6 by using *A. thaliana* wild type (Col-0), as shown in Supplementary Figure S1. This figure suggests that the results of the targeted BS-Sanger are reproducible by using the SIMON method.

Very small changes of methylation rate often have significant biological significance (Vial-Pradel et al. 2018). We conducted power analyses of the Fisher's exact test in order to estimate the necessary number of samples (DNA-fragments) to detect a difference of 10% methylation. We investigated statistical powers in a range from 0.05 to 0.95 for proportion and a range from 100 to 900 for sample size, as shown in Figure 5. This result

showed that approximately 1,000 samples (fragments) were necessary to detect a difference of 10% methylation. In the case of our analyses, comprehensiveness was not necessary for the targeted methylation sites, but high coverage was necessary to detect small changes in the methylation rate. Therefore, we proposed a novel method, whereby we used in silico size sieving of DNA-fragments and primer matchings instead of whole-fragment alignment for the assignment of loci. The algorithm based in silico size sieving of DNA-fragments and primer matchings was approximately 10-fold faster than the conventional BS mapping algorithm (QuasR) (Gaidatzis et al. 2015).

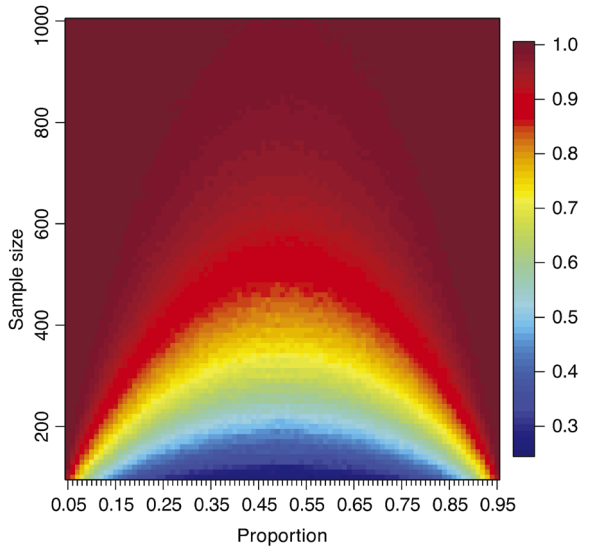

Figure 5. Power analysis of various sample sizes and proportions for Fisher's exact test.

We can estimate the cost benefit of this method when we want to detect 10% methylation changes and make a comparison with detection by the targeted BS-Sanger (see Table 1). As shown in Figure 5, 1,000 fragments/locus/samples are required. With the 4 bp barcode that we designed, up to 24 samples can be analyzed at once, and if we contemplate having 40 loci of interest to examine, it means that 960,000 fragments are necessary. In our experiment, we obtained a final yield of 3,885,305 fragments/9,737,840 raw read pairs (=0.3989904). Therefore, in order to obtain 960,000 fragments, 2,406,073 raw read pairs would be required (=960,000/0.3989904). Furthermore, 1 run of MiSeq Reagent Kit v3 (300 bp PE) (207,750 yen) generates 23,500,000 raw read pairs. 10 experiments could thus be conducted with 1 run where 2,406,073×10 is nearly equal to 23,500,000 raw read pairs. The cost would then become 207,750+10×9,523=302,980 yen for 10 experiments, or 30,298 yen per experiment. If we tried to use the targeted BS-Sanger, it would be preposterous to attempt to sequence 960,000 fragments. Instead of 1,000 fragments/locus/samples, it would be more reasonable to consider 10 fragments/locus/samples, which would provide preliminary information about the methylation levels. This means that instead of 960,000 fragments, we would be satisfied with only 9,600 sequenced fragments. If we suppose a yield of 100%, then 100 plates with 96-wells would be required to sequence all of these fragments.

Table 1. Cost benefit comparison.

Item	SIMON method	Targeted BS-Sanger
Number of samples	24	24
Number of loci	40	40
Required coverage	1000 fragments/samples/loci	10 sequences/samples/loci
Detection power	10% of variation	Preliminary information
Required total number of fragments/sequences	2,406,073	9,600
Number of fragments in 1 run	23,500,000	96
Number of runs required	1/10	100
Number of primer sets required for PCR amplification	960	40
Workflow after bisulfite treatment	Performing 960 PCR reactions with different primer sets (10 96-well plates), Mixing PCR fragments together, Purifying the mixture, Sending the sample mixture to the laboratory that performs the NGS.	Performing 960 PCR reactions with 40 different primer sets (10 96-well plates), Ligation in 960 plasmid vectors, Transformation of 960 batches of *E. coli*, Isolation of clones, Colony PCR (More than 9,600 reactions in order to obtain 10 positives clones/locus/ samples), Amplification of positive plasmids, Purification of the plasmids, Preparation of the samples for sequencing, Sending the plates to the laboratory that performs the sequencing.
Estimated minimum time required to complete the preparation after PCR	1–2 days	1 week
Cost of 1 run (excluding workflow materials)	302,980 JPY (for 10 experiments)	38,400 JPY
Comparative cost for the experiment	30,298 JPY (1/10 of a run)	3,840,000 JPY (100 runs)

The prices vary on the market, but a typical price for a full 96 well-plate sequencing is 38,400 yen (TaKaRa Bio). Therefore, the experiment described here, which involves analyzing the methylation levels of 24 samples at 40 different loci, should cost at least 3,840,000 yen when using the targeted BS-Sanger, and the data generated would be many times less accurate (1,000 fragments vs 10 fragments). Our assumption that the other specific expenses are at least comparable is not negligible when compared with the cost of the sequencing (barcode-extended primers, competent cells, other reagents, and materials). Taken together, the amount of lab work required for a similar study carried out by the different methods, once again, favors our new method.

When coverage is an important factor, such as for studies of methylation levels, and when the research focuses on many samples and many loci, the SIMON method has demonstrated its superiority when compared with the targeted BS-Sanger. Furthermore, if it is used to its full potential with several independent experiments that include dozens of samples and dozens of loci, SIMON should still be a very cost-effective alternative, even when other solutions exist.

In conclusion, we demonstrated that the SIMON method is capable of efficiently measuring small variations in patterns with precision and at a single nucleotide resolution. This method, which combines pre-existing techniques of PCR amplification and NGS, presents cost-effective advantages for in-depth studies that focus on pre-identified loci. The workflow and detection power are both significantly improved when compared with other techniques, such as targeted BS-Sanger. It is particularly efficient for studies that involve several samples and that especially focus on several specific loci, rather than whole-genome analysis. Owing to the depth of the analysis, small variations can be precisely determined, even with large numbers of samples and loci.

Acknowledgements

The authors are grateful to Ms. Yamakawa and Mr. Suzuki for their helpful technical support. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [grant numbers JP18H03330, JP18K06297, JP26291056]; The Ministry of Education, Culture, Sports, Science and Technology (MEXT) KAKENHI [grant numbers JP17H05659, JP16H01246, JP26114703].

Disclosures

The authors have no conflicts of interest to declare.

References

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* 27: 361–368

Baubec T, Akalin A (2016) Genome-wide analysis of DNA methylation patterns by high-throughput sequencing. In: Aransay A, Lavín Trueba J (eds) *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*. Springer, Cham, pp 197–221

Bewick AJ, Ji L, Niederhuth CE, Willing EM, Hofmeister BT, Shi X, Wang L, Lu Z, Rohr NA, Hartwig B, et al. (2016) On the origin and evolutionary consequences of gene body DNA methylation. *Proc Natl Acad Sci USA* 113: 9111–9116

Bird AP (1978) Use of restriction enzymes to study eukaryotic DNA methylation: II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *J Mol Biol* 118: 49–60

Bonnefoix T, Bonnefoix P, Callanan M, Verdiel P, Sotto JJ (2001) Graphical representation of a generalized linear model-based statistical test estimating the fit of the single-hit Poisson model to limiting dilution assays. *J Immunol* 167: 5725–5730

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* 452: 215–219

Compere SJ, Palmiter RD (1981) DNA methylation controls the inducibility of the mouse metallothionein-I gene lymphoid cells. *Cell* 25: 233–240

Clark SJ, Harrison J, Paul CL, Frommer M (1994) High sensitivity mapping of methylated cytosines. *Nucleic Acids Res* 22: 2990–2997

Charif D, Lobry JR (2007) SeqinR 1.0-2: A contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M (eds) *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations*. Springer Verlag, Berlin, pp 207–232

Gaidatzis D, Lerch A, Hahne F, Stadler MB (2015) QuasR: quantification and annotation of short reads in R. *Bioinformatics* 31: 1130–1132

Gonzalgo ML, Jones PA (1997) Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). *Nucleic Acids Res* 25: 2529–2531

Hill JT, Demarest BL, Bisgrove BW, Su YC, Smith M, Yost HJ (2014) Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. *Dev Dyn* 243: 1632–1636

Holliday R, Pugh JE (1975) DNA modification mechanisms and gene activity during development. *Science* 187: 226–232

Ishibashi N, Machida C, Machida Y (2013) ASYMMETRIC LEAVES2 and FASCIATA2 cooperatively regulate the formation of leaf adaxial–abaxial polarity in *Arabidopsis thaliana*. *Plant Biotechnol* 30: 411–415

Iwasaki M, Takahashi H, Iwakawa H, Nakagawa A, Ishikawa T, Tanaka H, Matsumura Y, Pekker I, Eshed Y, Vial-Pradel S, et al. (2013) Dual regulation of *ETTIN* (*ARF3*) gene expression by AS1–AS2, which maintains the DNA methylation level, is involved in stabilization of leaf adaxial–abaxial partitioning in *Arabidopsis*. *Development* 140: 1958–1969

Jee J, Rasouly A, Shamovsky I, Akivis Y, Steinman SR, Mishra B, Nudler E (2016) Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. *Nature* 534: 693–696

Kojima S, Iwasaki M, Takahashi H, Imai T, Matsumura Y, Fleury D,

Lijsebettens MV, Machida Y, Machida C (2011) ASYMMETRIC LEAVES2 and elongator, a histone acetyltransferase complex, mediate the establishment of polarity in leaves of *Arabidopsis thaliana*. *Plant Cell Physiol* 52: 1259–1273

Lindsay H, Burger A, Biyong B, Felker A, Hess C, Zaugg J, Chiavacci E, Anders C, Jinek M, Mosimann C, et al. (2016) CrispRVariants charts the mutation spectrum of genome engineering experiments. *Nat Biotechnol* 34: 701–702

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133: 523–536

Magoč T, Salzberg SL (2011) FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27: 2957–2963

Matsumura Y, Ohbayashi I, Takahashi H, Kojima S, Ishibashi N, Keta S, Nakagawa A, Hayashi R, Saéz-Vásquez J, Echeverria M, et al. (2016) A genetic link between epigenetic repressor AS1–AS2 and a putative small subunit processome in leaf polarity establishment of *Arabidopsis*. *Biol Open* 5: 942–954

Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 33: 5868–5877

Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R (2009) ShortRead: A bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 25: 2607–2608

Phillips T (2008) The role of methylation in gene expression. *Nature Education* 1: 116

Rand K, Qu W, Ho T, Clark SJ, Molloy P (2002) Conversion-specific detection of DNA methylation using real-time polymerase chain reaction (ConLight-MSP) to avoid false positives. *Methods* 27: 114–120

Semiarti E, Ueno Y, Tsukaya H, Iwakawa H, Machida C, Machida Y (2001) The ASYMMETRIC LEAVES2 gene of *Arabidopsis thaliana* regulates formation of a symmetric lamina, establishment of venation and repression of meristem-related homeobox genes in leaves. *Development* 128: 1771–1783

Suzuki M, Jing Q, Lia D, Pascual M, McLellan A, Greally JM (2010) Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol* 11: R36

Suzuki MM, Bird A (2008) DNA methylation landscapes: Provocative insights from epigenomics. *Nat Rev Genet* 9: 465–476

Takahashi T, Matsuhara S, Abe M, Komeda Y (2002) Disruption of a DNA topoisomerase I gene affects morphogenesis in Arabidopsis. *Plant Cell* 14: 2085–2093

Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, Abaan HO, Albert TJ, Margulies EH, Green ED, et al., NISC Comparative Sequencing Program (2010) Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res* 20: 1420–1431

Vial-Pradel S, Keta S, Nomoto M, Luo L, Takahashi H, Suzuki M, Yokoyama Y, Sasabe M, Kojima S, Tada Y, et al. (2018) *Arabidopsis* zinc-finger-like protein ASYMMETRIC LEAVES2 (AS2) and two nucleolar proteins maintain gene body DNA methylation in the leaf polarity gene *ETTIN* (*ARF3*). *Plant Cell Physiol* 59: 1385–1397

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* 37: 853–862

Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A (2017) Strategies for analyzing bisulfite sequencing data. *J Biotechnol* 261: 105–115

Xiong Z, Laird PW (1997) COBRA: A sensitive and quantitative DNA methylation assay. *Nucleic Acids Res* 25: 2532–2534